NHR FAU

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

Energy Efficiency for HPC & AI

The Second Free Lunch Is Over. The Free Joules Are Not.

Four Illusions.



Ayesha Afzal

Erlangen National High Performance Computing Center (NHR@FAU)

HPC Roundtable Munich · 27 October 2026

What the chips do with the 97%

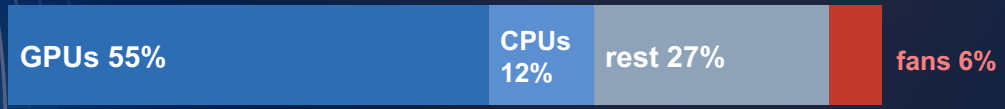
1.03

PUE of SuperMUC-NG (LRZ): 97 of every 100 W reach the IT

Helma (NHR@FAU): ~1.03–1.05 (air cooling: ~1.6–1.8)

The building is solved. This talk is about the 97%.

GPU node (NHR@FAU Alex measurement via ipmi-sensors)



Where does the 97% go? 33% never reaches a GPU or CPU

Source: D. Kranzlmüller (LRZ), HPC Roundtable 2025

Why now: demand, law, scale – and power wall

THE STAKES

945 TWh

Demand (IEA)

data centres by 2030

~2× 485 TWh (2025)

PUE ≤ 1.2

Law (EnEfG)

new German data centres from Jul 2026

waste-heat reuse 10 → 20 %

75–100k

Scale (Gigafactories)

accelerators per EU AI gigafactory site

bids due 3 Dec 2026

300 → 1800 W

Chip power

per NVIDIA GPU: V100 → Rubin (reported)

AMD: MI250X 560 W → MI355X 1400 W

10 MW

Fixed MW budget

planned new NHR@FAU data centre

Helma (~700 kW): 768 H100/H200 GPUs

The law regulates the building. Nothing yet regulates how the IT uses its 97 %.
The gigafactory call asks for ”best practices in energy efficiency," without a metric for the IT.

Under a fixed megawatt budget, energy efficiency is throughput: **every watt saved is a watt for more science.**

Sources: International Energy Agency, Energy and AI (2026); German Energy Efficiency Act §11, requirements via TÜV Rheinland; EuroHPC AI Gigafactories call (2026), AMD/NVIDIA data sheets (V100 SXM2 300 W, HGX B300 1,100 W); Rubin ~1.8 kW Max-Q as reported (M.-C. Kuo 2026; up to 2.3 kW Max-P)

The free lunch is over, again

2005

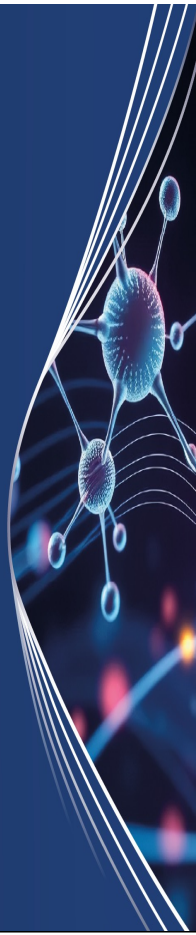
Clock speed stops scaling

Sutter, "The Free Lunch Is Over": software must go parallel to get faster

2023

Energy efficiency stops scaling

FP64 and bytes per watt stall. What still scales must be earned: by number format and by operating point.



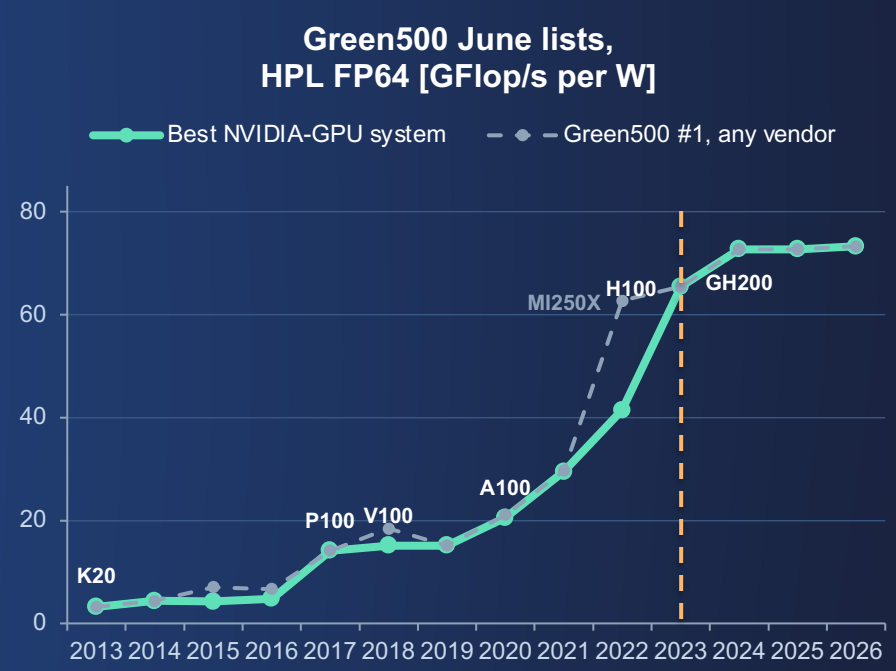
The second free lunch is over. The free joules are not.

Free joules: a joule you can save for free, without losing performance.

Sutter, "The Free Lunch Is Over", Dr. Dobb's Journal, March 2005. Evidence for 2023: next three slides.

The free lunch was efficiency – and it ended in 2023

EVIDENCE 1 OF 3 · MEASURED, WHOLE SYSTEMS



Not an NVIDIA artefact

the overall #1 flattened too:
62.7 (2022, AMD) → 73.3 GF/W (2026)

Blackwell at #91 (A100-era level)

best B200 system: 31.5 GF/W
– Green500 #91 (Nano4-b)

Not a blip

top 10 unchanged since Nov 2025
– a first in Green500 history

Fair caveat

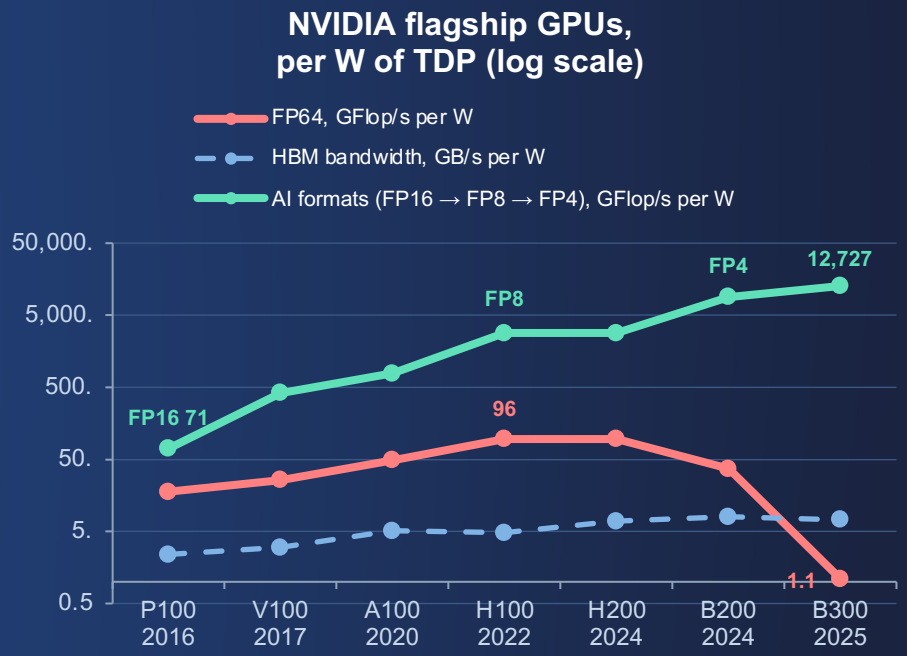
HPL is FP64; Blackwell traded FP64 for AI formats – next slide

2013–23: buy new hardware, get +35 %/yr. **Since 2023: +4 %/yr.**

Sources: TOP500.org Green500 June lists 2013–2026; HPCwire, 2 Jul 2026.

Where did the efficiency go? Into AI formats

EVIDENCE 2 OF 3 · CHIP DATA SHEETS, PEAK PER W AT TDP



×174

AI-format peak per W, P100 → B300

96 → 37 → 1.1

FP64 GF/W:
H100 → B200 → B300
(AMD: MI300X 218 → MI355X 56)

~7 GB/s per W

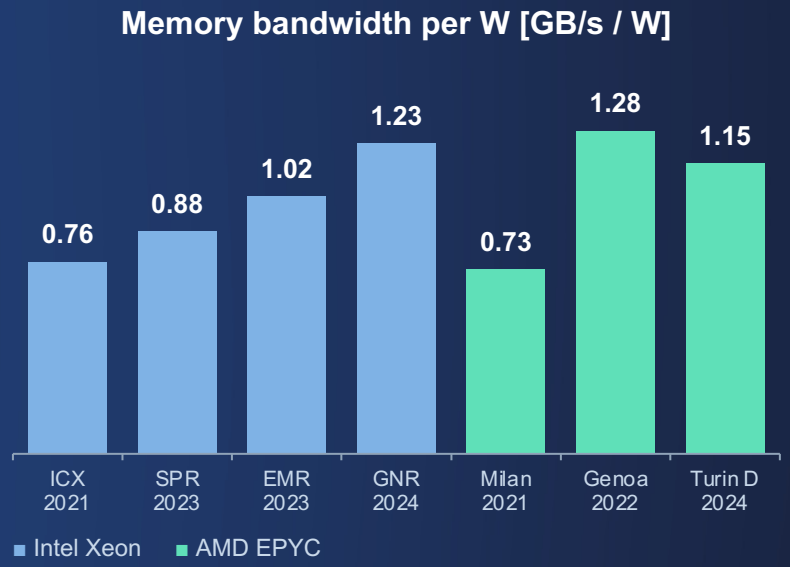
bytes per W flat since H200 – what memory-bound codes consume

On our FP64 science, the hardware dividend is gone.

Sources: Vendor data sheets (dense peak, best of vector/matrix FP64; SXM TDP). Transistors kept coming (15 B → 208 B), but Blackwell stayed on a 4 nm-class node. The new watts bought FP4, not FP64, not bytes. B300 FP4 ≈ 14 PF dense (spec varies by source).

CPUs: bytes per watt come from memory, not silicon

EVIDENCE 3 OF 3 · RECENT INTEL & AMD SERVER CPUs



What moved bytes per watt

+16 %	**ICX → SPR** DDR4 → DDR5, 8 channels
+21 %	**EMR → GNR** 8 → 12 channels, DDR5-6400
+75 %	**Milan → Genoa** DDR4 → DDR5 and 8 → 12 channels
−10 %	**Genoa → Turin D** same memory, 2× cores, +39 % TDP

Every gain is a memory step. AMD since 2019 (Rome → Turin): only ~5 %/yr.

Sources: data sheets, channels × MT/s × 8 B ÷ TDP. ICX 8380, SPR 8480+, EMR 8592+, GNR 6980P (DDR5-6400), Milan 7763, Genoa 9654, Turin 9965 (DDR5-6000 official; at 6400: 1.23), Rome 7742 (0.91). Peak, not sustained.

Two kinds of joules – the map of this talk

Of every joule, ask: **does it buy science?**

Free joules
buy no performance → remove everywhere

Traded joules
buy speed → choose

1 ILLUSION 1

“Energy is a hardware number”

hides in

the bottleneck and the model

2 ILLUSION 2

“Power capping saves energy”

hides in

a cap that never binds

3 ILLUSION 3

“One recipe fits every chip”

hides in

one equation

4 ILLUSION 4

“Faster is greener”

hides in

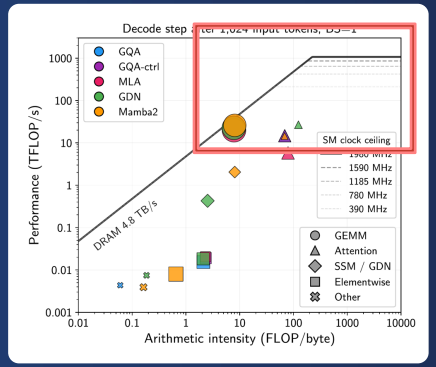
peak instead of lifetime work

Three of four illusions hide free joules: **most of the saving needs no trade-off at all.**

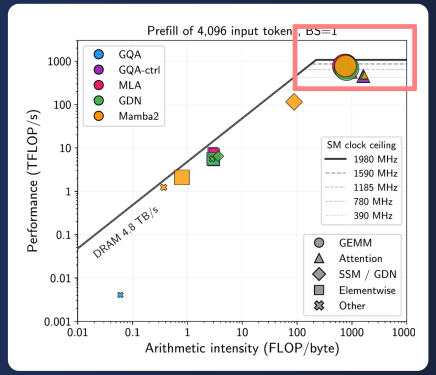
Illusion 1: energy is a hardware number

FREE JOULES · THE WORKLOAD DECIDES, NOT THE CHIP

① Same chip, different bottleneck

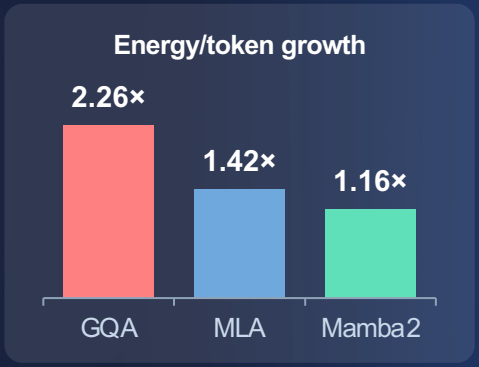


Decode: memory-bound.
Lower SM clock → same tokens/s, less energy.



Prefill: compute-bound.
Lower SM clock → proportionally slower.

② Same tokens, different model architecture



4K → 16K tokens of context, H200.
KV-cache traffic sets the slope.

-24% … −32%
decode energy with SM clock locked at 780 MHz

< 1%
throughput loss, all five architectures

0.5 MW
50 W saved × 10,000 GPUs, continuously

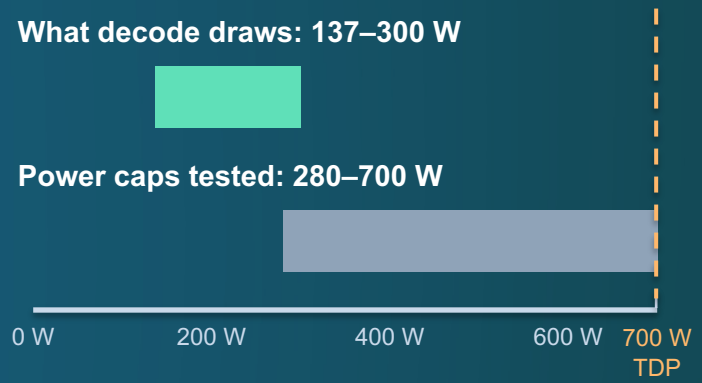
Same chip, same tokens – **the bottleneck and the model set the energy.**

Source: Ma et al.: PPAM 2026 (LLMs: arXiv:2605.11999). Nsight Compute per-kernel roofline, H200, vLLM.

Illusion 2: power capping saves energy

FREE JOULES · UNDER A KNOB THAT NEVER BINDS

AI: LLM decode · H200



The cap is never reached:
Clock locking wins in every configuration.

HPC: molecular dynamics / stress case · H200

5 of 11

AMBER benchmarks never reach the power envelope

≤ 440 W

GROMACS all 6 benchmarks, at any cap

≈100 vs 80 GF/W

FIRESTARTER benchmark on H200:
Clock locking wins over power cap

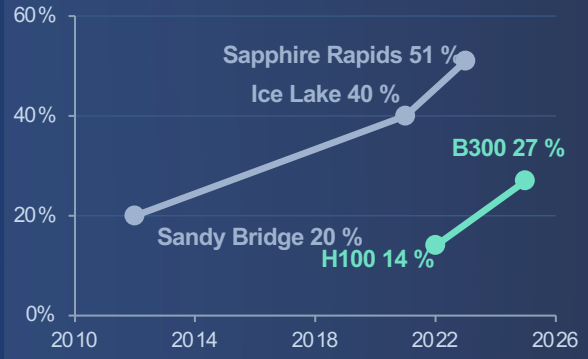
A power cap limits the watts you were not using.
A clock limit removes the watts you were wasting.

Ma et al., PPAM 2026; Afzal et al., PPAM 2026; Afzal et al., Euro-Par Wksp 2025

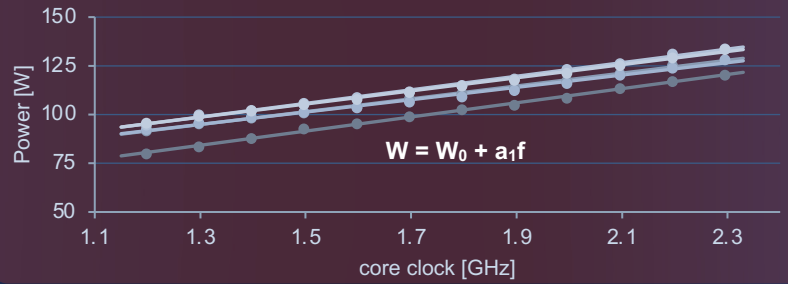
Illusion 3: one recipe fits every chip

FREE JOULES · ONE EQUATION, TWO ANSWERS

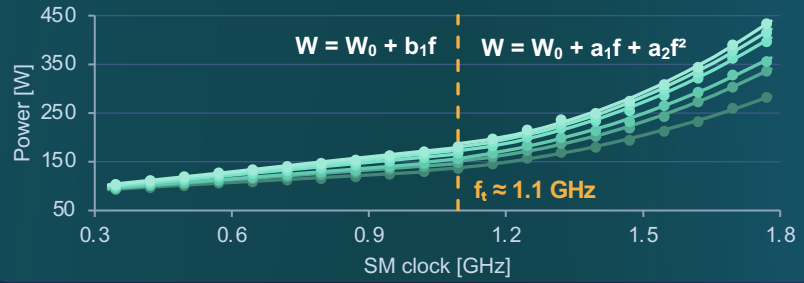
Baseline power W_0, share of TDP



CPU: race to idle wins (top clock) Ice Lake · 5 AMBER inputs



GPU: race to idle fails (knee f_t) H100 · 6 GROMACS inputs



One energy-optimal clock from CPU: 2015
MSc thesis → GPU: 2026 PPAM Best Paper

$$f_{opt} \approx \sqrt{\frac{W_0}{a_2}}$$

big W_0 → run fast
steep a_2 → slow down

W_0 and a_2 decide: same equation, opposite answers.

Sources: Afzal, MSc thesis FAU 2015; Afzal et al., PMBS@SC23, IJHPCA 2025, PPAM 2026 (Best Paper). CPU: LIKWID, GPU: nvidia-smi.

Illusion 4: faster is greener

TRADED JOULES · PEAK VS. LIFETIME WORK

73 GF/W

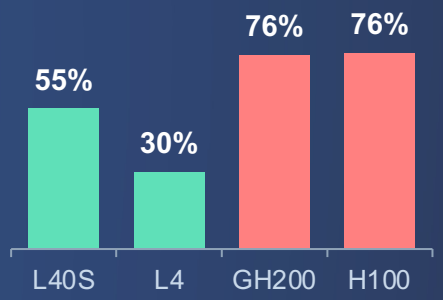
The belief: buy the Green500 #1 (GH200) – HPL in FP64

→

The test: €10M budget · GROMACS 6 benchmarks · 5-year lifetime – which GPU produces the most science?

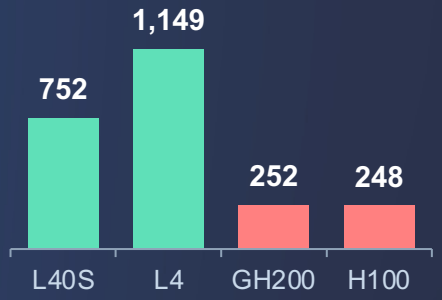
1 Where the money goes

GPU price, share of 5-year TCO



2 GPUs the same €10M buys

number of GPUs



3 3.2–3.8×

more **work per TCO** for L4 / L40S vs. H100/GH200

4 The flip

Low multi-GPU efficiency:
the L4 falls from top to last place

The fastest GPU is not automatically the greenest – nor is the slowest.
Only a model (performance · power · TCO) tells you which.

Source: Wattlytics web platform, wattlytics.netlify.app, Afzal et al., arXiv:2604.08182. Green500 June 2026.

Where the free joules go next

OUTLOOK · TWO HYPOTHESES · ONE RESEARCH UNIT

H1 The GPU sweet spot drifts up

Newer GPUs: smaller a_2; baselines rising → higher f_{opt}, smaller clock dividend – GPUs drift toward CPUs

H2 Free joules move from kernels to the gaps between them

Asynchrony becomes an energy design principle: stragglers waste up to 30 % of LLM training energy; synchronous training swings AI-factory power by tens–hundreds of MW

Mod4Comp · DFG Research Unit FOR 5880 · FAU · TU Dresden · FZ Jülich · 2026–2030

Measure	Model	Simulate	Decide
node power, NVML, LIKWID	Roofline/ECM + W_0, a_2, f_t per device	runtime & energy of parallel MPI programs	clock, cap, scheduling, procurement

Sub-project (PIs: A. Afzal, G. Wellein): the parallel simulator – carrying the energy models to emerging hardware.

Goal: **predict the energy before running the job or buying – or building – the hardware.**

Sources: Afzal et al., PPAM 2026; Chung et al., Perseus, SOSP 2024; Microsoft et al. 2025 (power swings).

Don't expect to always buy efficiency. Choose it.!!!

Harvest

Centres & operators

make the efficient operating point the default:
lock clocks per job or phase; caps only where they bind

Count

Everyone who reports

joules per job next to GPU-hours,
with a protocol

Choose

Procurement, funders, vendors

the operating point by what binds (GPUs, MW, €)
score work per joule at the sweet spot, not peak FLOPS

The second free lunch is over. The free joules are not.

Over to the AI Factories Roundtable

HANDOVER · ROUNDTABLE “AI FACTORIES”

5 MW

50 W per GPU × 100,000 accelerators – continuously

1. **Efficiency is an operating point. Who chooses it for a 100,000-GPU factory – user, scheduler, or firmware?**
2. **Should AI-factory tenders score work per joule and report joules per job?**
3. **Will vendors publish the energy datasheet, like idle power, threshold clock, and what a cap actually throttles – the way they publish peak FLOPS today?**

Thank you

Questions welcome

ayesha.afzal@fau.de

Papers behind this talk

GPUs

Modeling and Chasing the Energy-Efficiency Sweet Spots in Modern GPUs.
A. Afzal, M. M. Li, M. Panzlaff. PPAM 2026, **Best Paper Award**. doi:10.48550/arXiv.2607.00819

The Illusion of Power Capping in LLM Decode.
B. Ma, A. Afzal, J. Eitzinger, G. Wellein. PPAM 2026. doi:10.48550/arXiv.2605.11999

GROMACS Unplugged: How Power Capping and Frequency Shapes Performance on GPUs.
A. Afzal, A. Kahler, G. Hager, G. Wellein. Euro-Par Workshops 2025. doi:10.48550/arXiv.2510.06902

Clusters

Wattlytics: A Web Platform for Co-Optimizing Performance, Energy, and TCO in HPC Clusters.
A. Afzal, G. Hager, G. Wellein. 2026. doi:10.48550/arXiv.2604.08182

CPUs

Analytic roofline modeling and energy analysis of LULESH.
A. Afzal, G. Hager, G. Wellein. IJHPCA 2025. doi:10.1177/10943420251363711

SPEChpc 2021 on Ice Lake and Sapphire Rapids InfiniBand Clusters.
A. Afzal, G. Hager, G. Wellein. SC Workshops 2023. doi:10.1145/3624062.3624197

The Cost of Computation: Metrics and Models for Multicore Systems.
A. Afzal. Master thesis, FAU, 2015. doi:10.13140/RG.2.2.35954.25283

Funding & partners

DFG Deutsche Forschungsgemeinschaft

NHR FAU